2

RL-TR-92-244
In-House Report
October 1992

ROME LABORATORY

ADVANCED INFORMATION PROCESSING

John M. Pirog

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Rome Laboratory
Air Force Materiel Command
Griffiss Air Force Base, New York

This report has been reviewed by the Rome Laboratory Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RL-TR-92-244 has been reviewed and is approved for publication.

APPROVED: John R Bratge

JOHN R. BRATGE, Chief
Intelligence Data Handling Division

FOR THE COMMANDER: Garry W. Barringer

GARRY W. BARRINGER
Technical Director
Intelligence and Reconnaissance Directorate

If your address has changed or if you wish to be removed from the Rome Laboratory mailing list, or if the addressee is no longer employed by your organization, please notify RL (IRDS) Griffiss AFB NY 13441-4114 This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)

2. REPORT DATE
October 1992

3. REPORT TYPE AND DATES COVERED
In-House

4. TITLE AND SUBTITLE
ADVANCED INFORMATION PROCESSING

5. FUNDING NUMBERS
PE - 62702F
PR - 4594
TA - II
WU - PL

6. AUTHOR(S)
John M. Pirog

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Rome Laboratory (IRDS)
32 Hangar Road
Griffiss AFB NY 13441-4114

8. PERFORMING ORGANIZATION REPORT NUMBER
RL-TR-92-244

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Rome Laboratory (IRDS)
32 Hangar Road
Griffiss AFB NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Rome Laboratory Project Engineer: John M. Pirog/IRDS/(315)330-3222

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Current text processing techinques fall short of required accuracies. Text retrieval and filtering techniques are still largely based on keywords. Several new techniques and their combinations, can provide far greater accuracy in text retrieval and dissemination. Many of these techniques are briefly explored and their integration into the real world is also discussed.

14. SUBJECT TERMS
Text Information, Text Handling, Information Dissemination, Text Filtering.

15. NUMBER OF PAGES
36

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Figures

APPENDICES

Executive Summary

Information is a very powerful weapon. It can achieve its full potential only if it can be gathered and accessed in a timely, efficient manner. Improvements in collection make the job of gathering easy. In many cases its so easy that too much information is gathered. On the other side, accessing this information has never been easy and is becoming more difficult as more information is gathered. Current operational systems barely meet the needs of the typical day to day information analyst. Difficulty in retrieving just the information they want and nothing else is hampered by old methods, such as keyword matching. The system can be separated into two different entities (1) information retrieval (retrospective search) and (2) information filtering (dissemination). Differences are slight and for the purposes of this paper, no distinction is made. Most problems found in the use of profiles (a description of the information needed, best known as a list of keywords) is the inversely proportional attributes of recall and precision. Increasing one, most always decreases the other for any given technique. Seven different profiles are described as basic extensions to the basic keyword profile. The keyword profile with boolean logic is currently the most popular due to ease of implementation. When a weighting scheme is added to this profile it becomes a "weighted profile". These have the benefit of introducing "human" knowledge via the weights into the information processing process. Though much better than keyword alone techniques, they are difficult to set up. Proximity profiles are profiles that use word combination frequencies to eliminate some of the ambiguity of the English language. These encompass such things as phrases and acronyms. A sentence profile is a step up from this in that it takes the grammical structure of a sentence to imply something about what words might be used. A sentence usually has a single idea and this is exploited via sentence profiles. Profiles that utilize the location of text in a message are called format profiles. These utilize subject, body, and heading type information for retrieval and filtering. Finally, all of the profiles mentioned have their advantages and disadvantages, but together (or used in combination) they increase precision.

Another technique for bridging the gap between man and text is the use of transformation algorithms to transform text into a form more easily processed. These transformations are called text signatures. Here, text is transformed into a simplified representation of the original. Two techniques, signature analysis and

N-grams, are the ones undergoing the most research today. Both methods produce a list of entities (not always words) that describe that piece of text. Hopefully, each piece of text will have a different signature. What will the future for information processing hold? A concept of mutually assisted information retrieval and filtering. In this future system, the system will assist in the construction of profiles, utilizing examples of what the user wants to see. This will be combined with well known clustering techniques to ensure the most accurate information profiles possible. Achieving this "future" will require some experimentation in the combination of profiles and the use of text transformation and user interfaces to the system. Only after experimentation can one hope to realize the full potential of these techniques and the full potential of information's power.

This paper discusses the current state of information processing (retrieval and filtering) within the Air Force. It also describes current research on extensions to the technique currently implemented. Most of the discussion focuses on the usefulness and problems with profiles for information retrieval and filtering. Recall and precision issues are discussed as are some advanced techniques for translating text into a more useable (by computer processes) form. Finally, integration in the operational Intelligence Data Handling System is discussed along with a concept for future information processing.

Chapter 1

Background

1.1 Purpose

The purpose of this report is to provide a background on current information processing techniques; discuss the motivating forces behind information processing; provide a glimpse of what may be coming down the pike in terms of advanced techniques; and finally, to provide a view of the future in terms of its relationship to the Intelligence Data Handling System. While there are many different kinds of information (e.g. imagery, text, voice, sensor), it is the textual information that this paper will focus on. It is the intent of the techniques presented not to understand text, but to retrieve it with high recall and precision.

1.2 Motivation

In todays information age, information is becoming as important a commodity as oil or gold. Information today is more than just a collection of knowledge, it is a very powerful weapon not only in industry, but in the military as well. Knowing this, people have been, and will continue to collect as much information as possible about almost everything in the world. Evidence of this can be seen daily, more magazines, more papers, more television stations, 24 hour a day news broadcasts, and probably the largest source of information - computer networks. Improvements in communication systems have resulted in cheap accessibility to almost any piece of electronic data in the world, from almost anyplace in the world. This has resulted in a glut of electronically distributed textual information (the message). So much information is now being gathered that the user of this information, the Information Analyst, can not utilize all of it. In fact, the average Information Analyst often does not have time to read even the smallest amount of information they have filtered out of this glut. This means that much of the information collected is "wasted". It is the utilization of this information, and the power that is derived from it, that is the primary reason information is collected in the first place.

1.3 Current Process

In order to cope with the increasing volume of information, the analyst must be capable of rapidly turning their information needs into a request that results in obtaining the information they want and only that information. Many systems exist today that provide some tools or mechanisms to deal with these requests. These automated systems filter out the requested information. Unfortunately, these systems do a poor job at either filtering or translating their request into a filter. In either case, too much information, not enough information or the wrong kind of information is provided to the analyst. So the problem is not a lack of information, but instead, poor access to that information. There exists two distinct, yet very similar, types of information access systems. They are the **Information Retrieval (IR)** systems and the **Information Filter (IF)** systems. Information Retrieval (also known as retrospective search) is access to data bases of information via some sort of algorithm(s). This is where information is already stored in a database (and possibly indexed) and then pulled out via some sort of query mechanism. Information filter is information access to newly arriving information via a filter that sifts desired information from all that is arriving. This is where information is arriving and only those desired pieces of information should be displayed to the user. A message dissemination (mail) system is a good example. Automated systems currently exist for both types of information access, but both fall short in terms of accuracy (a combination of recall and precision), maintenance, and complexity.

Current IR systems require that the analyst know the type, scope and sizes of information in the data base, the structure of the data base itself and the "language" of the data base, in order to access historical information. Unfortunately the analyst is rarely able to articulate information access requirements with precision considering the amount of knowledge about the data base the analyst is required to know. This results in queries that often leave unanswered questions as well as new questions. Two IR metrics, recall and precision, have been shown in many cases (depending on the techniques used) to be inversely proportional, as you increase one the other goes down. This is the case because in order to maximize recall, the query must be defined to ensure all possibly related information is accessed. This results in accessing more than desired information in order to ensure nothing is missed. Conversely, if precision is maximized, an exact description of what one desires must be specified, but something inevitability does not get specified and information is missed.

Recall and Precision

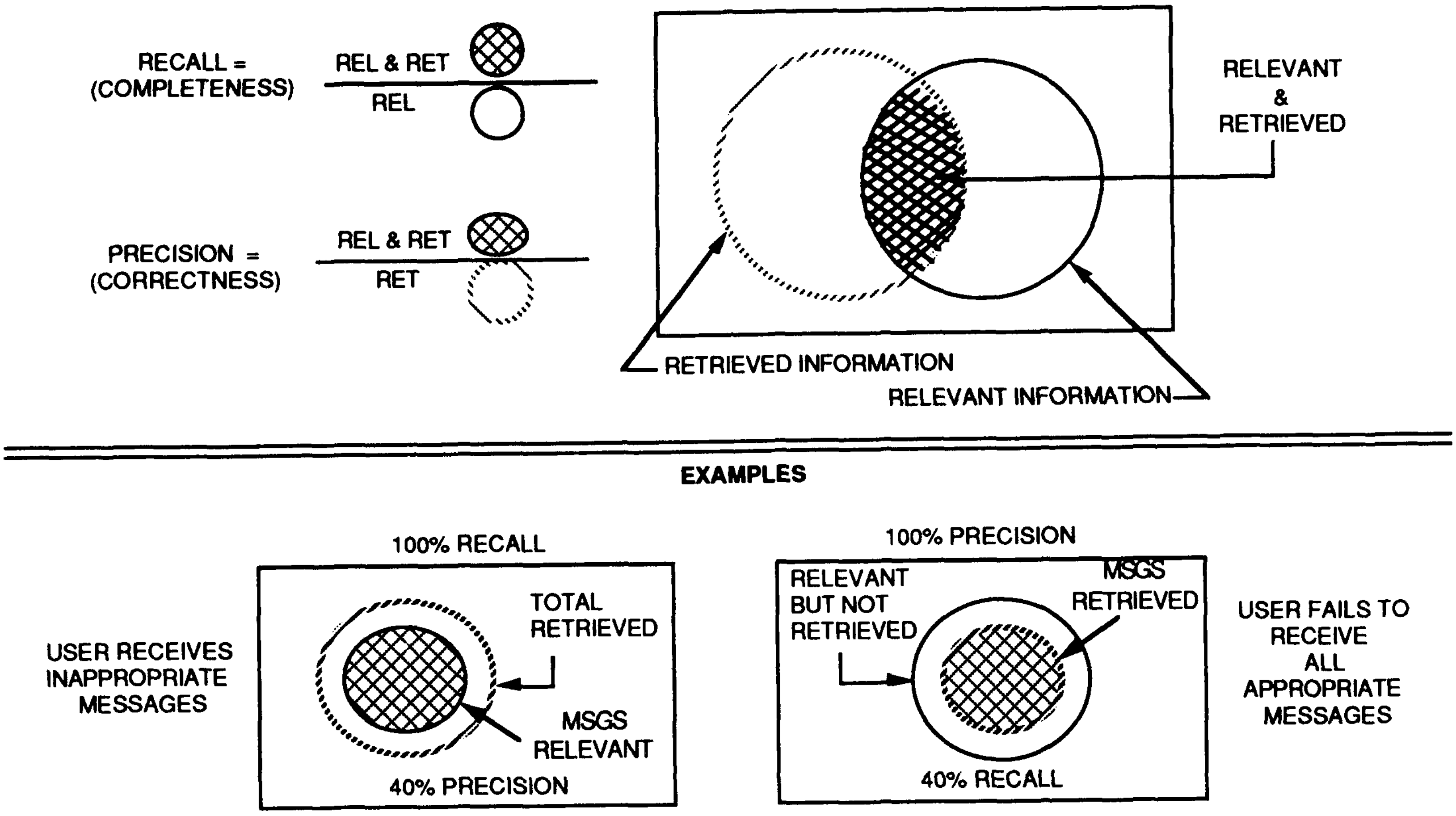


Figure 1

Current IF systems require that the analyst have knowledge about what kinds of information is going to be arriving and then describe the ideas, concepts or relationships desired in an "information filter" also known as a profile. A profile (in a keyword sense) is simply a list of words. Often the keywords can be found in boolean combination for example:

space **AND** craft **AND** (foreign **OR** soviet **OR** russian .....)

The analyst must manually create these profiles using a list of known keywords that somehow describe their desired interest areas. These profiles are then matched against incoming text. When any text arrives that matches any of the profile, its called a "hit" and sent to a user's mailbox. Which mailbox(es) receive this message is dependent upon who subscribes to that profile. Some users can have many profiles and hence many mailboxes. While these Information retrieval and filter systems work for the most part, they do have their share of problems.

1.4 Current Problems

There are many existing IF and IR systems today. In the area of IF systems, there are a couple large ones currently in use throughout the military services. The one with the largest user base is the Modular Architecture for the Exchange of Information (MAXI). This system uses hardcoded profiles consisting of keywords linked together with boolean operators. The MAXI is routinely operating at peak capacity, which is enough to swamp an analyst. MAXI is scheduled to be replaced by a system called the DoDIIS Automated Message Handling System (AHMS). The throughput of the AMH component of the DoDIIS is estimated at a peak volume of 16,750 incoming messages a day, or almost 50 million characters every 24 hours (ESD-ICP-A100, DEC89). Obviously, an inordinate amount of manpower would be needed to accurately analyze and interpret this information in its bulk form. Profiles are used in this system as the way to convey analyst information needs and to perform the filtering function. Construction of a MAXI profile is a slow, manual task, and prone to errors. This manual process results in the analyst spending too much time creating these "information profiles" and little time performing the analysis of the data. When a "hit" occurs, the text is sent to mailboxes as specified by the profile distribution list. Users subscribe to a profile so they can receive information. Many times they pick a profile that is close enough to what they are looking for because it is easier (and faster) than building one by scratch. The DoDIIS AMHS is improving upon the keyword aspect of IF systems by coming up with the idea of constructing "concepts". Concepts are keywords related to one another in a hierarchical fashion. Weights are attached to the words as an added measure of relative importance. Building these concept profiles are not much easier than the MAXI keyword boolean profiles and have the added complexity of weighting. Assignment of weights can make or break the advantage of a concept profile.

Hierarchical Structure of a Concept

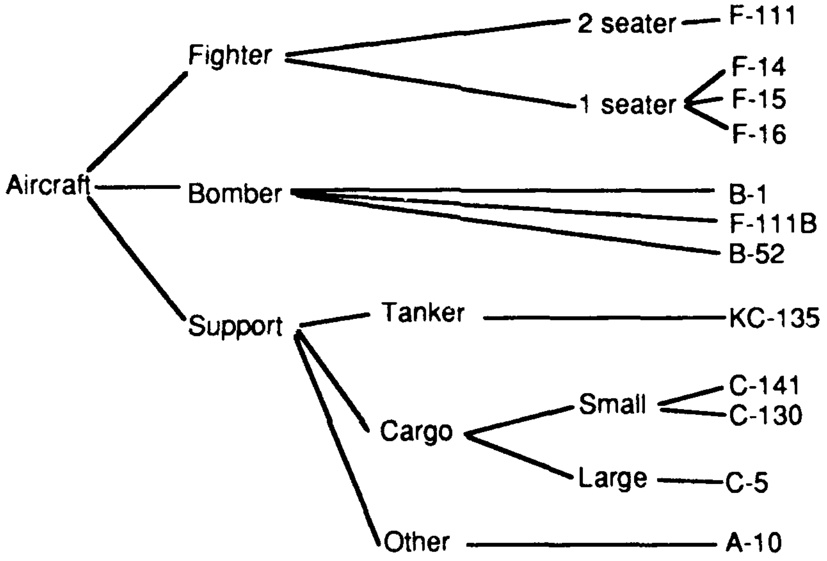


Figure 2

Information retrieval systems have long been a commercial product, although they are almost entirely keyword based with connectives like boolean logic. Examples include TRW fast data finder, Proximity by Proximity, and FERRET. When using these systems, the analyst must quickly formulate their ideas into a query language. While these "small" profiles (the query) are saved for use again, they often fall far short of what is needed and result in the need to generate another one. On many occasions, accessed information leads the analyst to begin to browse the database in search of answers to their questions. With each return of information from a query, another query is formed to follow up on some line of reasoning. It is these Ad-Hoc queries that have become the norm when doing exhaustive searches. Since these systems provide little more than a query language, the user is left to guessing how to best (efficiently) query the system. Smarter queries will lead to faster, more accurate answers to the analyst's questions. Making the user smarter at performing queries or building profiles will help alleviate these problems but is not practical. Too much user time would be spent creating these queries and profiles, leaving little time for analysis. Making the system itself smarter, is an easier, more cost effective solution. The next section deals with new techniques that might help make a smarter information retrieval or filter system.

Chapter 2

Technologies for Text Representation

2.1 Technologies and Their Promise

There is now a current plethora of techniques and technologies that are being developed to support better Information Processing. The techniques receiving the most amount of research are those basic techniques exploiting the various types of profiles. The basic keyword profile, being the most popular and best understood, yet one of the poorest. More advanced profiles (which are basically derivations of the keyword profile) such as boolean, physical proximity, etc. offer more promise. While each of these approaches have their advantages and disadvantages, their combination offers a greater potential. Lesser known techniques that show promise such as n-grams, signature analysis, message via clustering and the newcomer, neural networks, are now just emerging . The following techniques apply equally to both text filtering (IF) and text retrieval (IR) systems. No distinction is made between the two types of systems to make the discussion more readable.

2.1 Text Profiles

The most common and well understood method for filtering and retrieving of free text uses a profile (a query being a very small profile). In this technique a list of related words (or phrases) is used by some sort of comparison algorithm. As the complexity of the profile increases, the recall and precision increase, and the time taken to generate the profile increases. Most commercially available retrieval methods categorize text based on the presence or absence of words or phrases in boolean combination. Some methods provide weighting for keyword emphasis and physical proximity for phrase and compound recognition. The addition of some natural language structures or semantic knowledge provide even greater power to these techniques. However, the maintenance of these profile structures can be a difficult problem, if hundreds of profiles exist on the system simultaneously. Maintenance of profiles is necessary since as the world changes, new words, phrases or structures are used to describe current events.

2.2.1 Keyword

The keyword profile represents the allowable letter combinations (words) within the domain of interest. These words are typically organized in a list of some sort. The words may or may not have the same meaning causing some errors to be introduced (precision) when using this method. A good example is the word "fire". The various meanings can be found in Figure 3.

Definitions of FIRE

Fire - to dismiss a person from employment
Fire - a destructive burning
Fire - to have an explosive charge at the right time
Fire - to utter with force and rapidity
Fire - to apply fire or fuel to
Fire - to throw with speed

Figure 3

The keyword profile can be expanded by utilizing phonetic as well as thesaurus information to expand the keyword list. This is done because, often words are misspelled or not every word to describe an information need can be thought of. For example:

Keyword: Kadify	Phonetic words: Quadfi, Kadiffi, Qadify......
Keyword: Fire	Thesaurus expansion: Volley, sniping, salvo.....

The problem with this type of profile is that any message with any occurrence of a word from the profile will match providing a very low precision performance (receiving messages containing one of your words although it has nothing to do with what you want). Recall performance, one would suspect to be very good (won't miss a thing) but instead is not. This is due to the inexactness of developing the profile. Unless every known word, and their phonetic variations and synonyms are expressed for a certain need, an idea indirectly related to the need, will be missed. This type of profile is developed without regard to the types of information that is in the data base.

2.2.2 Boolean Profiles

Boolean keyword profiles are nothing more than the logic needed to connect the keywords together. For example, consider the following boolean expression :

Small **AND** (Airplane **OR** Airbus) **AND** (foreign **OR** 3rd world)

This profiling technique is an improvement over simple keywords, however, recall and precision performance is still low. Recall performance is better, but still low, since the boolean expression is too mechanical. For example, if a message arrives with the words airplane and small but nothing about countries, then this message does not match the expression even though it probably is about the same event. However, precision is improved from a simple keyword profile for the logical associations among the keywords provide a simplistic view of information content. This content view allows many of the different concepts for a particular word to be tossed out of consideration. An example of this would be the following boolean profile:

fire **AND** gun **AND** target **OR** animal

This profile would consider only the following definitions of fire (from lengthy list in Figure 3).

Fire - to have an explosive charge at the right time
Fire - to throw with speed

The greatest problem associated with boolean profiles is their complexity. A typical profile might contain 300 keywords put in logical association with one another. The precise creation of this sort of profile is nearly impossible at best.

2.2.3 Weighted Profiles

Application of relevancy or weights to a profile can assist in the discrimination of information. Its use is not limited to only certain types of profiles. Weights can be applied to any type of profile whether it is keyword, expression, sentence or any combination. In a weighted keyword profile the weight indicates the word's importance to the desired information. A weight of 1.0 for any word or expression indicates the word perfectly identifies the information requested. Conversely, a weight of zero implies no association with the requested information. Care must be used when assigning weights since improperly assigned weights can cause a profile to deviate from the intended information requirement. For example consider the following weighted keyword profile:

San Salvador	1.0
arrest	0.8
El Salvador	1.0
threat	0.6
drugs	0.8
illegal	0.4

Here, the profile covers anything concerning El Salvador and San Salvador and the flow of illegal drugs between them. This profile should accept messages about the two countries and anything of a illegal, threatening, drug nature. This will occur only if both countries are mentioned since their weights are set to 1.0. If either country is left out of a message, that message might not be retrieved. A 1.0 weight in a keyword profile conveys the idea that the words El Salvador and San Salvador are the most important to the profile. Another attribute of weighted profiles is that the frequency of each of the words found in the data base will affect the recall and precision. In the above example, if the word drug was found with great frequency (such as a pharmaceutical data base), then its weight of .8 will cause a large portion of the data base to be considered, as that word is found often. Either a lesser weight or a more discriminating choice of a word would help this situation. Most users of weighted profiles do not take any of this into account when generating profiles, hence their poor performance.

2.2.4 Physical Proximity Profiles (Expressions)

Physical proximity profiles are formed by the combination of frequently used words in a sequence. These profiles are based on groups of words typically associated with each other. Examples include the Untied States of America, National Aeronautics and Space Administration, or President Bush. One can exploit these type of word groupings and eliminate the individual keyword influences. Utilizing this information can increase precision by further filtering out undesired information. An examination of the actual message text can probably uncover the compound phrase "National Aeronautics and Space Administration" is used with some frequency. Obviously, this phrase represents one concept and these words should not be profiled independently. A statistical computation of association between words can be performed to derive the compound phrases. The recall and precision of this type of approach is a solid improvement over using only keywords but is limited to only compound phrase related information. The reason for this improvement is that physical proximity profiles profile some word definition context, allowing other definitions to be eliminated.

2.2.5 Sentence Profiles

A sentence, as defined by Webster's Dictionary, is "a grammatically self-contained speech unit consisting of a word or syntactically related group of words that expresses an assertion, a question, a command, a wish, or an exclamation". Sentences can be located by identifying words between end punctuation. Two approaches exist for identifying word grammatical usage, (1) statistical and (2) linguistic. The linguistic approach employs an English textbook form of analysis in order to identify the grammatical structures. The problem with this type of approach is that it generally requires extensive semantic knowledge in the form of dictionary entries for the specific domain under analysis. This process is better known as Natural Language Processing and currently is not practical for large bodies of text. Statistical methods for generating the use of a word exist but their accuracy is poor. They function by tagging words with parts of speech attributes. These "tokenizers", utilize a "pre-tagged" corpus of text, similar to the body of text being tagged. This method will achieve approximately 85-90% correct tagging for that type of text. The accuracy of tagged text outside the pre-tagged text will be much lower. Sentence profiles and their tagged parts of speech provide a means to limit the meanings of words.

The recall and precision of sentence profiles is considered to be high. The reason is that these profiles are considered to possess some grammatical association or contextual information. Their use is limited though, by the inadequate ability to properly tag text.

2.2.6 Synonym Modifications to Profiles

No two people produce text in the same way, even if they are writing about the exact same thing. This is due to variations associated with synonyms, spelling variations, word morphology, transliterations, and acronyms. Although many electronic thesauruses exist, accurately determining how and when to use synonyms is critical. For example, consider the word "launch" in a space domain and it's synonyms originate, start, set going, thrust, fire off, eject, propel, drive, motor boat etc. None of these synonyms would normally be used inside the space domain text with the possible exception of "thrust" or "fire off", but neither are good synonyms for launch in this case. Therefore, using any of these words as a replacement for "launch" in a space domain profile could create a situation where large amounts of irrelevant text would be retrieved. The use of synonyms provides a greater robustness to the sentence profile. This is because the expansion of the word list using synonyms and spelling variations increases the recall. Yet caution must be used so as to keep the irrelevant documents to a minimum (decrease precision).

2.2.7 Format Profiles

Format profiles can impart some added meaning via the location of keywords within a body of text. Most all documents have some basic rules governing their construction. Examples are: headers, subject lines, From/to blocks and the introduction, main body and conclusion paragraphs. The identification of these structures is very useful in determining the types of information that is typically found in these areas. Similar words found to be in the same areas of different messages increases the probability that the messages are about the same subject. The problem is that formats can vary widely. The process of extracting message structure information and normalizing it is called zoning. Zoning is based on the location of special words, characters and indentation. Fairly simple to implement with a high degree of accuracy, zoning can provide improvements in recall without decreasing precision.

2.3 Profile Combinations

The profile "types" described previously in this document represent several levels each having positive and negative recall and precision attributes when used to represent information requirements. Since no single profile representation by itself completely describes all concepts and ideas, a combination of profiles would seem to offer the best overall representation. Starting with the basic keyword profile, each additional profile used in conjunction with the basic keyword will improve recall. Precision can suffer though, due to the recall combining (more hits), which will impact precision. Most profiles are robust enough to work within most any text domain. Two profiles require manual manipulation (weighted keyword, boolean) or are limited in their scope (sentence profiles). Combined profile performance would depend on the text being investigated. Figure 4 is a summary of the types of profiles and their benefits;

Types Of Profiles

PROFILE TYPE	DESCRIPTION	RECALL/PRECISION
KEYWORD	LIST OF WORDS	LOW/LOW
BOOLEAN KEYWORD	WORDS WITH LOGICAL OPERATORS	LOW/HIGH HIGH/LOW
WEIGHTED KEYWORD	WORDS WITH IMPORTANCE ATTACHED	HIGH/MEDIUM
COMPOUND PHRASE	FREQUENTLY COMBINED WORDS (e.g. UNITED STATES)	MEDIUM/LOW
SENTENCE	WORDS WITH WORD USAGE	MEDIUM/HIGH
SYNONYM MODIFICATIONS	WORDS EXPANDED WITH ALTERNATE WORDS	HIGH/LOW
FORMAT	LOCATIONS OF WORDS AND THEIR MEANING	LOW/MEDIUM

Figure 4

2.4 Text Signatures

A signature is a pattern that uniquely identifies a single message. Unlike profiles which are information requirement descriptions that can span many different small concepts or even one large one, message signatures are the representation of a single message and hence, can sometimes be only a partial idea with many "holes" (How often does a single piece of text meet all the information requirements?). The signature representation of the message provides a means to cluster messages based on a comparison of their respective text signatures. The "closer" signatures are to one another, the more likely they are about the same subject. Most forms of text signatures are based on the occurrence and physical proximity of alphanumeric strings or words within the text.

2.4.1 Signature Analysis

Signature analysis is the process of developing a signature representation of a piece of text. Developing a signature representation begins with the removal of the common words. These common words are also called stop words. Stop words are removed to reduce the size of signatures describing the text and to eliminate meaningless words such as : "on", "the", "it", and "for". All punctuation is also removed. Adverbs and adjectives can also be removed from the text before the signature is produced, although the current school of thought is to leave them in. Other words that need to be removed require a little more thought. If the database of messages are domain specific, then there may be words that appear too frequently to be discriminating. For example, if major portions of the database were about United States agriculture, the word United, States, agriculture, farm, etc., may have to be removed since these words may not differentiate the various subtopics in the database. It is possible for signature analysis to go astray and produce different text signatures for similar texts. This problem stems from the following facts:

- **Stop Words**: Stop word list should be created statistically for each database. If the stop word list is not maintained, eventually performance will decrease as the database contents changes over time.

- **Recall and Precision**: These values are inversely proportional for keywords. Since this technique is rooted in the keywords found in the text, similar recall and precision problems can appear. Adding additional profile information (parts of speech tagging, weights etc) produces more precise signatures.

- **Syntactics**: Most signature analysis techniques do not evaluate word order . For example, ... on the Earth.... does not have the same meaning as .... earth on the .... even though the same words are used and the signatures will be similar.

2.4.2 N-grams

An N-gram is a member of a set defined by all N-tuples of length N formed over an alphabet of symbols. For instance, all 2-grams over English alphabet are contained in the set (aa ab ..... az, ba ..... zz). An N-gram signature is formed by collecting the rarest N-grams occurring in the words in the text. The two character N-grams (2-grams) for the word "duck" are "du", "uc", and "ck". One of the 3-grams in duck is "uck" which also occurs in "lucky", "pluck", and "tuck", etc. Since "uck" occurs too frequently, it would not be made part of the signature. Based on this least frequent N-gram approach, a distinctive signature is formed using this technique since only text which has many of the same words could have similar N-grams. N-grams does not differentiate the importance that might be associated with different words.

2.4.3 Neural Networks

The pattern comparisons that are performed in the process of profiling and retrieval are quite similar to the capabilities provided by Neural Network technology. Neural networks possesses the ability to learn patterns and to match this pattern to new patterns in constant time. They also possess the ability to differentiate (classify) between two seemingly similar entities. It is this later ability that most closely resembles statistical clustering of text. Unsupervised learning can potentially be used to extract distinguishing characteristics of text messages. Finally the ability of neural networks to deal with "fuzzy" or imprecise information (within the patterns) will increase recall, but could decrease precision. The graph shown in Figure 5, represents the capabilities of Neural Networks applied to the Information Processing problem.

Neural Capabilities compared to Information Processing Areas

(CONNIE DESIGN PLAN, 15 FEB 90)

MESSAGE HANDLING NEURAL NETWORKS	PROFILE	RETRIEVE	CLUSTER	FILTER
LEARNING				
ADAPTATION	1		2	
GENERALIZATION	●	●	●	
STATISTICAL CLUSTERING	3		●	
COMPRESSION	●	●	●	
RECOGNITION				
CLASSIFICATION		5		6
SPEED		●		●

1,3 - Adpatively create profiles for novel messages

2 - Adaptively develop clusters based on novel topics

4 - Associate for synonym replacement

5,6 - Associate for conceptual search

Figure 5

Chapter 3

Advanced Information Processing

3.1 Integration into the IDHS

Technology is advancing at an extremely fast rate. It is unfortunate that the operational world is usually several years behind state of the art in computer technology. Even though the operational world is catching up by fielding equipment that is state of the art, it is often fielded as a standalone component. Millions of dollars have been invested in the current systems and architectures and no one can afford to throw it all out and start over. It is this reason that the integration of advanced techniques and tools such as Information Processing, is not a simple "plug and play". New technology/techniques to be truly useful, must be integrated into the current architectures carefully. Two approaches are discussed below that can help to integrate the advanced information processing technology into the operational world as well as other technology.

First, if there isn't a place to put the new technology, make one. This concept arises out of the old school of thought to develop an interface between any two entities and they will be guaranteed to work. The problem with this approach is that one gets locked into the interface and any change to either component causes change to one or both of the other components. Instead, one should develop a generic interface component, that opens up the architecture to the outside. Half of this generic interface will be tightly coupled to the current architecture to ensure compatibility. While this approach may cost a lot more in the near term, in the long run it can save many times the cost. An example, currently under development, is a prototype interface component called Generic Intelligence Processor (GIPR). This component will allow the integration of advanced text processing components into an operational environment. The objective is to develop an architecture for the introduction of advanced text processing capabilities into the operational environment. This development will consist of two distinct components, a server and some application tools. The server portion will allow a user to change the interface between the environment and the new technology. The application portion will provide the support to the new technology itself. The GIPR design will allow for incremental insertion of advanced text processing capabilities, with little to no impact to the user. In fact, the ultimate goal is to increase the users productivity without the user even being aware of the productivity.

The second approach is to utilize whatever is existing and adapt the development of the advanced technique to that existing piece. This approach has already been used to integrate the initial GIPR in the existing message processing component. Message traffic is received through AUTODIN via the Communication Support Processor (CSP) and routed to Modular Architecture for the Exchange of Information (MAXI) for further dissemination. MAXI is used as a coarse filter and outputs the traffic via its RS-232 port to the GIPR prototype. This connection is a one way feed from MAXI to GIPR. Users still have access to all MAXI messages and their standard MAXI queues (integration without impact). Security and communication services are provided by the Workstation Support Environment (WSE) and Common User Baseline for the Intelligence Community (CUBIC) baseline products. The GIPR will continue to provide support with little modification when MAXI is replaced by the Automated Message Handling System (AMHS).

3.2 Information Processing Future

An entirely new method of information processing can be envisioned in light of the discussions above. This can be accomplished since the profiling schemes presented provide various representations of each piece of information. This information representation allows humans to define their information needs and for text to be represented by an alternate form (text signature). This helps to bridge the gap between a need for information and the information itself. Implementation of the techniques can be envisioned as below. Instead of a user developing a profile manually, then submitting it, both man and machine can work together to develop a profile (or combination profile). In a directed mode, the user guides the system in order to get just the right information. This is done by the user providing an example text message and the system developing a profile for that example. The system utilizes various algorithms to develop one or more of these profiles. As more examples are provided to the system, the more accurate the profile will become. This profile is more robust than a manually generated profile and more exact due to the profile being generated from the actual text itself. Signature analysis can also be used to translate text into another representational form that may be processed much faster. Of course this means that all incoming information would also have to be processed. This concept can be integrated with another mode of operating called undirected mode. Here, the system acts on its own to separate information into various areas (called clusters). Since their is no guidance by the user, the areas generated may not have any importance to the user. This latter technique is called clustering. Like information is grouped together, based upon the similarity of their text signatures. Combining the two concepts, one should be able to direct the clustering algorithm so as to cluster around

Information Flow

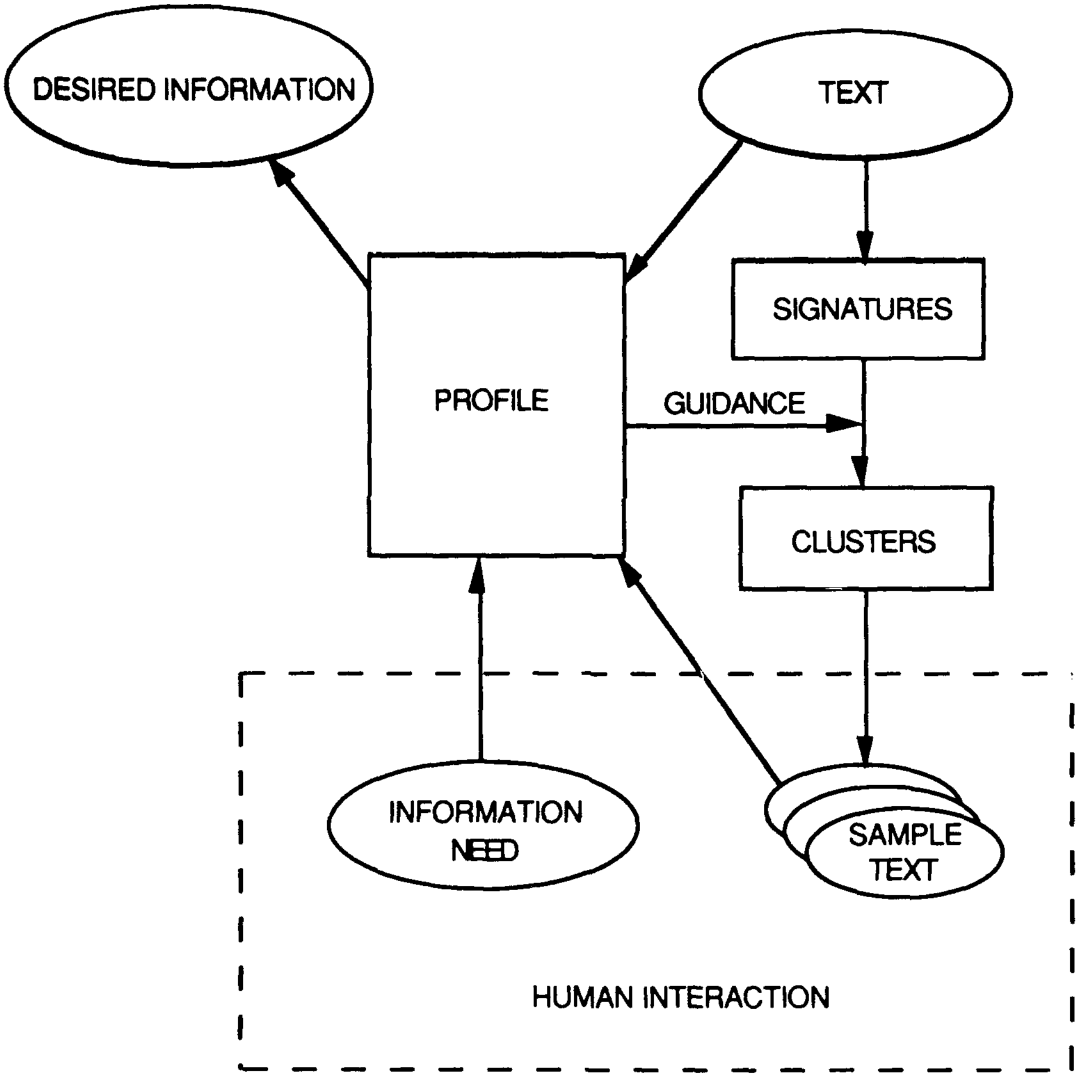
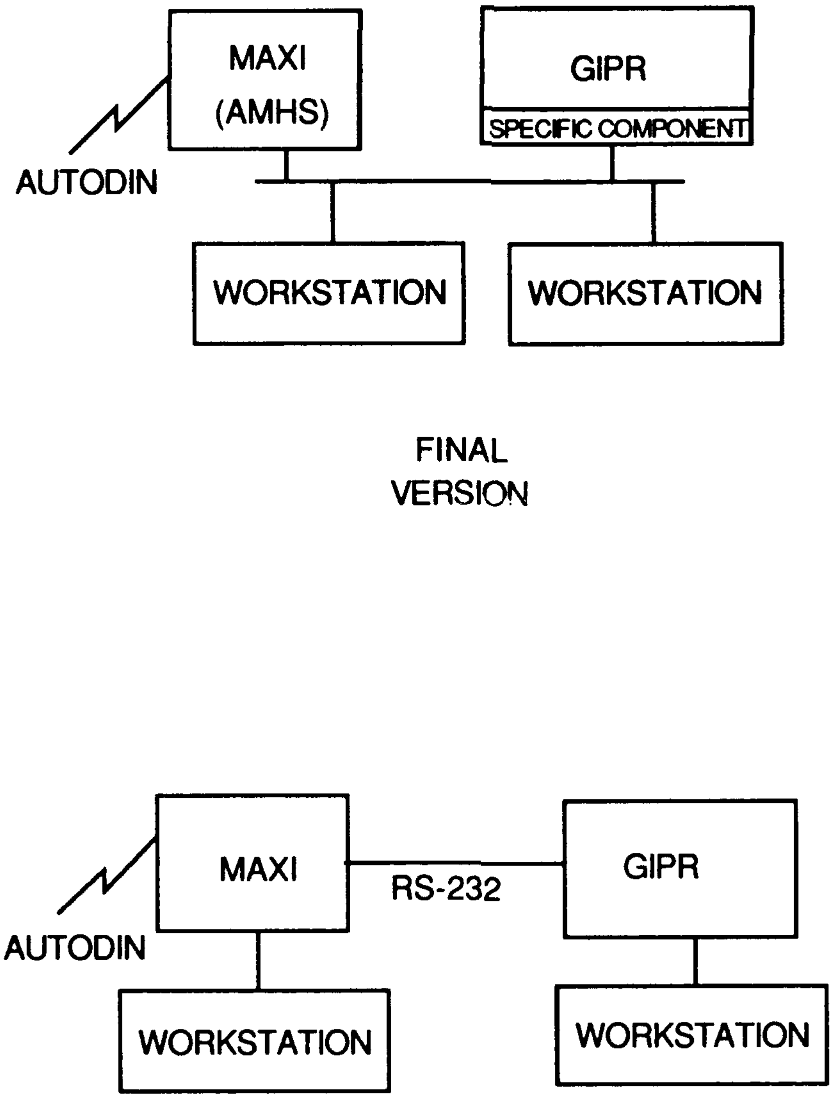


Figure 6

GIPR Integration



FINAL
VERSION

INITIAL
VERSION

Figure 7

areas of interest (profiles). This would also facilitate the addition of new example texts from the clustered sets in the profile. Finally, as with most techniques, accuracy will falter and clustering may bring in texts that have little to do with the desired area. In this case, the profile can have a "negative" example added to it. Here, the system is told it should not retrieve information similar to what's in this example. This provides even greater accuracy. All of these procedures can be repeated as many times as necessary in order to develop the best profile possible.

3.3 Summary

Information is a very powerful tool. Access to information provides the means to utilize that power. Problems with accessing information are well known and must be overcome or at least minimized. Presented in this paper were several techniques for translating information into a more usable (computer wise) form. The mapping of information needs into "profiles" is the result of the current way of performing Information Retrieval and Information Filtering. The translation of text into signatures is but another representation of information that is easier to process by computer algorithms. It is the common ground between text signatures and profiles that will allow the problems of information access to be naturally reduced without the need for extremely sophisticated processes and algorithms. There is no one solution to providing easy accurate access to information. Precision and recall issues tend to be inversely proportional no matter what the technique is used. It is only by the smart combination of various techniques presented here as well as others, that one can hope to increase accuracy of access.

3.4 Recommendations

Some functional text handling improvements could be made almost immediately by incorporating just some of the techniques described in this document into the existing Intelligence Data Handling System. The "new" profiling techniques would naturally integrate easier into the existing systems than the text signature techniques due to the profile techniques being nothing more than extensions to the existing profile concepts in use today. The creation of a single profile for simultaneous high recall and precision data base retrievals and the ability to browse a data base without having to learn a special query language could be realized. Continued research into the use of these profiles in combination within an operational environment (such as the IIPF) should be accomplished. Along with the research into profiles, effort should also be put forth on the development of text signatures. This relatively unexplored area holds much promise.

Appendix A

Glossary

Automatic Indexing
An automated technique for generating a list of keywords or other characterizations so that information may be easily obtained.

Cluster
A number of items related to one other by some common thread .

Compound
A group of words that co-occur with regularity and constitute a standard phrase (i.e., United States of America).

Directed Mode
Message grouping based on a given profile or text signature.

Expression
See compound.

Index
A list of words or characteristics which identify the information and location of information in a data base.

Information Analyst
Any person whose job is to transform and combine information into a form required by that individual or that individual's organization.

Lexical Analysis
The analysis of items of information in terms of their constituent morphemes, roots, words, acronyms, compounds, etc.

Message Cluster
A group of messages that have certain characteristics in common.

Precision
The proportion of the number of relevant messages to the total number of messages in the retrieval. EX: Have 200 messages, of which only 180 are relevant.

Profile
A group of characteristics that define a information request.

Recall
The proportion of the number of relevant messages accessed to the total number of relevant messages available in the message stream or data base. Everything is accessed that is desired but it comes with a lot of unwanted information.

Relevance Weight
A value calculated to show the importance of a single lexical item characterizing a message cluster.

Signature Analysis
The comparison of messages based on the similarity of their text characteristics. Text signature representations include N-grams, neural network representations, and transform imaging.

Stop Word List
A list of words that are unimportant or have little meaning to the task at hand (i.e. the, a, not, to).

Synonym
A word or expression accepted as a figurative or symbolic substitute for another word or expression.

Zoning
The process of identifying document structures in a normalized pattern or singular format.

Appendix B

Referenced Documents

1. Current Status of Automated Message Handling Systems, Technical Report, TR-RD-91-2, Planning Research Corporation, 1991.

2. Research and Development for Intelligence Data Handling, RL-TR-91-319, Rome Labs, October 1991.

3. Advanced Text Processing Concepts for Automatic Message Handling, Griffiss AFB, NY: RL/IRDP, August, 1988.

4. Topic User's Manual. Verity, Inc., Mountain View, CA: 1989

5. Topic System Administrator's Guide, Verity, Inc., Mountain View, CA: 1989

6. SACWARNS Advanced Textual Information Processing (Draft), Planning Research Corporation.

7. Connectionist Networks for Information Exploitation (CONNIE) Interface Design Document (Draft), Grumman Corporation, November 1989.

8. New Technology Database Generation: A Natural Language Message Understanding System Based on Cortical Through Theory, (Draft), Harris Corporation, 17 November 1989.

9. Connectionist Networks for Information Exploitation (CONNIE) Final Technical Report, Grumman Corporation, November1991.

10. Webster's Ninth New Collegiate Dictionary, Merrian Webster, 1985.

©U.S. GOVERNMENT PRINTING OFFICE

MISSION

OF

ROME LABORATORY

Rome Laboratory plans and executes an interdisciplinary program in research, development, test, and technology transition in support of Air Force Command, Control, Communications and Intelligence (C^3I) activities for all Air Force platforms. It also executes selected acquisition programs in several areas of expertise. Technical and engineering support within areas of competence is provided to ESD Program Offices (POs) and other ESD elements to perform effective acquisition of C^3I systems. In addition, Rome Laboratory's technology supports other AFSC Product Divisions, the Air Force user community, and other DOD and non-DOD agencies. Rome Laboratory maintains technical competence and research programs in areas including, but not limited to, communications, command and control, battle management, intelligence information processing, computational sciences and software producibility, wide area surveillance/sensors, signal processing, solid state sciences, photonics, electromagnetic technology, superconductivity, and electronic reliability/maintainability and testability.